Article

A Prototype to Investigate the Extent to Which Words with Specific Attributes Can Be Retrieved Using Granular Metadata

Liezl H. Ball
https://orcid.org/0000-0002-1483-0780
University of Pretoria, South Africa
liezl.ball@up.ac.za

Theo J.D. Bothma
https://orcid.org/0000-0001-7850-3263
University of Pretoria, South Africa
theo.bothma@up.ac.za

Abstract

Despite the growth in digital text collections, the ability to retrieve words or phrases with specific attributes is limited, for example, to retrieve words with a specific meaning within a specific section of a text. Many systems work with coarse bibliographic metadata. To enable fine-grained retrieval, it is necessary to encode texts with granular metadata. Sample texts were encoded with granular metadata. Five categories of metadata that can be used to capture additional data about texts were used, namely, morphological, syntactic, semantic, functional and bibliographic. A prototype was developed to parse the encoded texts and store the information in a database. The prototype was used to test the extent to which words or phrases with specific attributes could be retrieved. Retrieval on a detailed level was possible through the prototype. Retrieval using all five categories of metadata was demonstrated, as well as advanced searches using metadata from different categories in a single search. This article demonstrates that when granular metadata is used to encode texts, retrieval is improved. Relevant information can be selected, and irrelevant information can be excluded, even within a text.

Keywords: digital humanities; digital libraries; metadata; information retrieval; text collections; prototype

Mousaion
https://unisapressjournals.co.za/index.php/LIS
Volume 41 | Number 4 | 2023 | #14399 | 18 pages

https://doi.org/10.25159/0027-2639/14399
ISSN 2663-659X (Online), ISSN 0027-2639 (Print)

Introduction

Digitisation efforts from the past decades, together with the increasing number of born-digital items, have led to numerous digital collections. There are large-scale repositories, such as Google Books (https://books.google.com/), HathiTrust (https://www.hathitrust.org/) and the Internet Archive (https://archive.org/).

The availability of resources in digital form, together with the developments in technology, has opened up new possibilities for researchers. New methods and techniques are employed to explore texts and text collections, for example text and data mining, distant reading and visualisation (e.g., Michel et al. 2010; Nguyen et al. 2020; Senseney et al. 2021; Viiri 2014). Interesting studies using new techniques and methods have been done, for example, the investigation of cultural trends (Michel et al. 2010), the mapping of emotions in London (Heuser, Moretti, and Steiner 2016) or investigating gender in English fiction (Underwood, Bamman, and Lee 2018). Such studies can be considered as part of the developing field of digital humanities, an interdisciplinary field, where the humanities and technology overlap (Edmond and Lehmann 2021).

There is increasing interest to support researchers who wish to work with digital collections. This is evident in the development of tools and infrastructure that can be used to explore digital collections, for example, the Google Books Ngram Viewer (https://books.google.com/ngrams), a tool that shows the trends of words in the Google Books corpus, Voyant Tools (https://voyant-tools.org/), an online tool for the analysis of digital texts, and the HathiTrust Research Centre (https://analytics.hathitrust.org/), a platform for users to do computational analysis in a secure environment while complying with copyright restrictions (e.g., Jett et al. 2016).

Despite the development of various tools, there seems to be limitations in the way in which searching can be done currently. Researchers are often not only interested in items on a volume or book level, but in entities within volumes or books (Fenlon et al. 2014; Jett et al. 2016). These could be chapters, poems or pages (Fenlon et al. 2014). If users and researchers are able to pinpoint exactly what they wish to retrieve from a collection, research can be more exact. An example will be used to illustrate this point. A collection of essays, *The Oxford Book of American Essays*, was published in 1914. The editor of this collection is Brander Matthews. This volume contains essays by various authors, such as Benjamin Franklin, Edgar Allan Poe and Theodore Roosevelt, with different publication dates. Furthermore, some essays contain quotations from other authors. For example, Washington Irving quotes a poem by Drummond of Hawthornden (1585–1649) in one of his essays. If it is possible to retrieve on a fine-grained level, a user could specify that an essay from Benjamin Franklin should be included in his/her search even if it appears in a volume edited by another author, or, it could be possible to search for quotations of Drummond Hawthornden in work published after 1900. Not only would detailed bibliographic data be useful to create precise queries, other data about the texts and words in the texts could also be used to narrow down a search. For example, it could be useful to search for words with a specific meaning, or of a particular part-of-speech category. Furthermore, such data could be combined to create complex searches.

The authors of this article investigated the use of granular metadata to enable fine-grained retrieval in digital text collections. The limitation in current tools to search and filter on a detailed level in digital text collections has been reported on, and possible categories of metadata that could be used to improve retrieval have been identified (Ball 2020; Ball and Bothma 2022). The authors argue that by encoding texts with granular metadata, more advanced search options could become possible. This article reports on the implementation and testing of these suggestions and describes the development of a prototype that was used to demonstrate how retrieval can be improved when texts have been encoded with granular metadata.

Research Problem

It is evident that digital collections are increasing, and so is their use for research purposes. However, searching on a fine-grained level in text collections is limited, which restricts what researchers can do. This problem should be addressed and more should be done to enable researchers to retrieve specifically what is relevant to their need. This has led to the following question:

To what extent can words or phrases with specific properties be retrieved when texts are encoded with granular metadata?

In order to investigate this problem, a sample of texts was encoded with granular metadata and a prototype was developed. The prototype was used to test the retrieval of words or phrases with specific properties. The success of such a prototype can contribute significantly to the way in which researchers can explore digital text collections. Enhanced retrieval can enable researchers to ask very specific questions that are currently not possible, or at best, prohibitively cumbersome.

In the next section, some related work and projects will be discussed, after which the granular metadata that can be used to capture additional information about words and texts will be discussed. The rest of the article will be devoted to the discussion of the encoding of texts and the prototype.

Background and Related Literature

Various tools, projects, infrastructure and services have been developed to support researchers to use digital text collections effectively. Though the limitations in tools that engage with digital text collections have been discussed extensively (Ball 2020; Ball and Bothma 2022), a few examples and pertinent remarks should be given here.

The Google Books Ngram Viewer has already been mentioned as a tool used to study a large corpus and allows users to view trends of words over time; it has been used in various studies and inspired further research. Further features of this tool are that it allows users to search using some morphological or syntactic properties of words, for example, to search for all inflected forms of a word (e.g., *bring*, *bringing*, *brought*), to search for specific part-of-speech categories, or to search for instances where a word is dependent on another (e.g., where *lovely* modifies *place*) (Google Books Ngram Viewer n.d.). However, due to copyright restrictions, there are limited options to filter according to bibliographic metadata and retrieval is coarse.

The HathiTrust+Bookworm (https://bookworm.htrc.illinois.edu/develop/), on the other hand, has extensive bibliographic metadata, but limited other options, such as morphological and syntactic metadata. Some tools from the field of linguistics (e.g., the BNCweb: http://corpora.lancs.ac.uk/BNCweb/) have advanced search options if a user knows the relevant query language (Hoffmann and Evert 2006). There are also tools that are compatible with texts that have been encoded according to the Text Encoding Initiative (TEI) guidelines, which are widely used to represent texts in digital form (TEI n.d.), such as TXM (Heiden 2010). TXM is an example of a powerful but complex tool. Voyant Tools is another example of a tool developed for the analysis of texts, and it was designed to be user-friendly so that users who do not have specific knowledge about encoding or programming may use it (Welsh 2014). However, this tool also has limitations, such as the ability to search for words with a specific meaning or the relationship between words. More detailed descriptions of these tools are provided in Ball (2020). In cases where more detailed filtering and searching are possible, it typically requires programming expertise or knowledge of the underlying data (Ball and Bothma 2022).

One more detailed example will be given to illustrate the developments, as well as the limitations, in tools that allow users to search in and engage with digital text collections. Chronicling America (https://chroniclingamerica.loc.gov/) is a website that allows a user to search America’s historic newspaper pages from 1777 to 1963. It is developed by the National Digital Newspaper Program (NDNP), a partnership between the National Endowment for the Humanities (NEH) and the Library of Congress (LC).

The advanced search screen for Chronicling America has several interesting search options. This tool is a clear example of the attempts that are made to allow users to conduct an advanced search in a collection of digital texts. Users can filter according to basic bibliographic data, such as place of publication or language. Furthermore, the ability to search for certain words or a combination of words is allowed. This is done through the options that are similar to Boolean and proximity operators. The tool also has inbuilt language processing capability, automatically searching for variants of a word. For example, if a person searches for the word “strike,” instances of “strikes” are also returned. Importantly, the possibility of allowing users to search in certain sections of an information source is available in the feature that allows users to search in specific pages, such as the front page or a page number.

Despite the features that enable an advanced search in this tool, the way in which this could further be enhanced should be mentioned. An attempt is made to allow users to search in sections of the newspapers; however, this is limited to the level of a page. If this is extended to sections in the newspapers, and different sections, such as advertisements, cartoons, articles or letters, are identified, it could enhance retrieval even further. Instead of just filtering on a page level, a user can filter on sections. This is in line with the findings of Fenlon et al. (2014), where participants wanted to select entities within a text, and Underwood (2015), who discussed the need for the identification of genre of sections in a text. It has also been argued that useful tools to explore collections will benefit researchers in the field of digital humanities (Lansdall-Welfare and Cristianini 2020).

This example will be used to describe what could be possible if more data about the properties of words in a text are available. The results of the search for the word "strike" include instances with different meanings. There are instances where "strike" refers to the action of refusing to work as part of a protest, and there are instances where "strike" refers to the use in baseball. If the meaning of the word was known, the user could have been more precise in their search and only return items with a certain meaning, whereas currently all instances of "strike," regardless of meaning, are returned.

It is evident that tools and platforms to offer advanced search options are being developed. However, more research should be done to see to what extent further improvements are possible. In order to enable this type of enhanced searching, more data are needed about sections in the texts and the words and phrases in the texts. This additional data can be regarded as granular metadata.

There is much data about texts, and the components that make up texts, that can be useful. From the publication date of a whole work to the part-of-speech category of a word in a sentence, and various levels in between, can be considered. Possible metadata that can be useful for retrieval are evident in various tools and literature (e.g., Fenlon et al. 2014; Finlayson 2015; Lin et al. 2012; Underwood 2015). Five categories of metadata that can be used to capture additional data about texts have been identified and discussed in previous work (Ball 2020; Ball and Bothma 2022), namely, morphological, syntactic, semantic, functional and bibliographic. A brief summary is given for the sake of clarity. The morphological level relates to data about the word and the word structure, for example, the part-of-speech categories of a word. The syntactic level refers to the relationship between words, for example, the link between a subject and verb in a sentence. The semantic level refers to the meaning of a word. The functional level refers to structural features in a text, for example, chapters and footnotes that form different sections, or logical features, for example, a word used as a name of a person. The bibliographic level refers to data that can identify the text, such as title and publication date, but it has been argued that the bibliographic information should go beyond the volume level to also describe sections in a text, such as where one author quotes another (e.g., where Irving quotes Hawthornden, referred to earlier). The authors of this article argue that if information from these different categories can be used and combined, powerful and precise queries can be created.

In the next section, the prototype that was developed to enable fine-grained retrieval through using multiple categories of metadata will be discussed.

Method

Prototype

The ideas proposed by the authors of this article were tested using a prototype. A prototype is a preliminary version of a system (Suranto 2015). A prototype is not the final system but has some functionality. Prototypes are valuable for several reasons. A prototype can help to determine if a new idea is feasible, whether the concept works as intended and if the idea should be explored further before resources are wasted (Klimczak 2013). Through a prototype it is also possible to communicate a new idea (Klimczak 2013). There are different types of prototypes, ranging from systems that are very close to the final product to prototypes that are simpler and do not have as much functionality (Walker, Takayama, and Landay 2002). In prototype development there should be a balance between the amount of functionality and the speed with which the prototype can be developed (Klimczak 2013).

In order to test the extent to which specific words could be retrieved when granular metadata have been applied to texts, various steps are necessary.

- Specific elements in each category of metadata that are useful for retrieval should be identified and listed.
- The way in which these elements should be encoded needs to be established.
- A sample of texts should be encoded following the established encoding guidelines.
- A prototype that allows a user to search for words or phrases with specific properties should be built. This prototype should receive and process the encoded texts. The data should be stored in a database, and a user should query the database through an interface.

Specific elements for each category have been identified and suggested (Ball 2020). For example, it is suggested that on the morphological level the lemma of each word as well as the part-of-speech category of each word should be recorded. If the lemma of a word is known and encoded, it will enable retrieval of the lemma. The part-of-speech category will allow a user to specify which part-of-speech is relevant to his/her query. The way in which such metadata can be encoded has also been described. For example, it was suggested that each word is encoded in a word tag (<w>) and the lemma and part-of-speech category are encoded as attributes.

It is beyond the scope of this article to discuss all the elements and the encoding. The focus of this article is on the third and fourth steps, namely, the encoding of a sample of texts, and the development and testing of the prototype. However, some aspects of the elements and encoding will be mentioned in the discussion. More details may be found in Ball (2020).

Sample Texts

The data that were necessary to test this prototype are texts that were encoded with granular metadata. A prototype is by definition not a fully-fledged, production-ready system. It is a smaller version of the envisioned final product, with some of the functionality. Therefore, it is only necessary to have a dataset that is large enough to test the functionality of the system. If it does not work, then the idea can be abandoned; if it does work, then the idea can be explored further. As such, a sample of texts were selected for encoding. A purposive sampling approach was followed to select samples of texts to encode. The requirements for the selection of samples were that all the elements that were proposed were included in the texts. For example, it was suggested that direct speech should be encoded, as such there should be a sample of direct speech.

Six texts were selected for this project and are listed in Table I. Samples from the texts with relevant examples were selected and encoded. The table includes the following data: the title, the date the text was published, the author and the number of words from each text that was encoded.

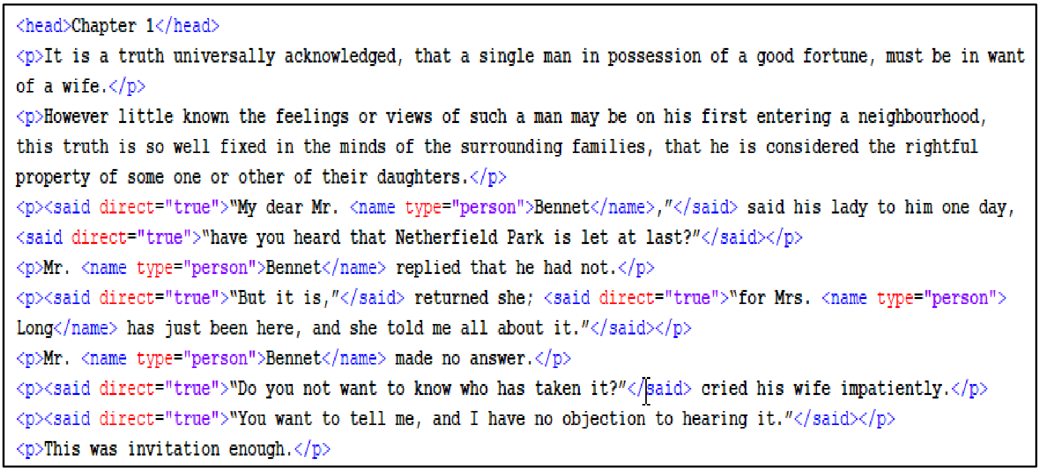
Table I: Texts selected for encoding

Nr	Title	Publication date	Author	Size (words)
1	*Pride and Prejudice*	1813	Jane Austen	468
2	*Middlemarch*	1817	George Eliot	733
3	*Ben-Hur: A Tale of the Christ*	1880	Lew Wallace	566
4	*The Life of St. Teresa of Jesus, of the Order of Our Lady of Carmel*	1904	Teresa of Ávila	901
5	*My Man Jeeves*	1919	P.G. Wodehouse	569
6	*Clouds of Witness*	1958	Dorothy L. Sayers	1119
Total				**4356**

The encoding of the various categories of metadata was split into two files. Although there would be some advantages to have all the encoding in one file, the volume of metadata is significant and as some encoding is done manually, it would hinder a human encoder. Each text then has two files associated with it. The one file contains the bibliographic and functional metadata and the other file the morphological, syntactic and semantic metadata. One advantage of this approach is that the encoding of different categories can happen in parallel.

The encoding process was partly manual and partly automated. Due to the labour-intensive process of encoding, it was essential to investigate the possibility of automating some of the encoding. The automated encoding of some categories of metadata is fairly well-developed. The Stanford CoreNLP (https://nlp.stanford.edu/software/) library was used for the tokenisation of the texts (identification of sentences, words and punctuation) and encoding on the morphological level (part-of-speech categories and the lemma) and the syntactic level (dependencies). Some manual corrections were necessary. There are no libraries for semantic encoding that are publicly available and reliable, and the semantic encoding was done manually. That is, the assignment of the meaning of the words with a WordNet label (http://wordnetweb.princeton.edu/perl/webwn) was done manually. The bibliographic and functional encoding was also done manually.

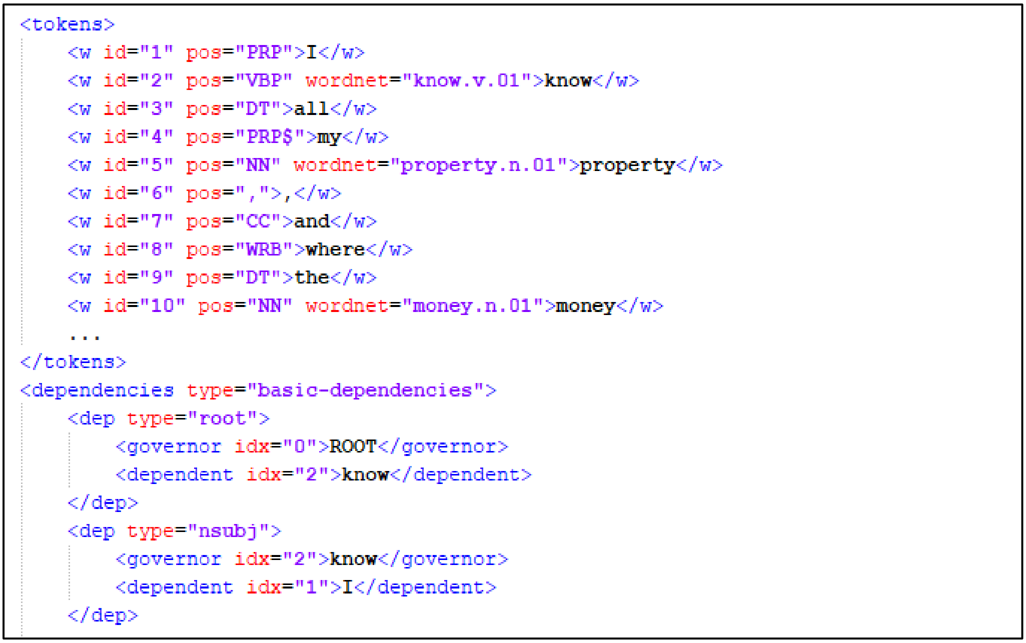
An example of the encoding of the functional metadata is shown in Figure 1. The heading of the chapter is encoded in head tags and the first paragraph is encoded in paragraph tags.



```
<head>Chapter 1</head>
<p>It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want
of a wife.</p>
<p>However little known the feelings or views of such a man may be on his first entering a neighbourhood,
this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful
property of some one or other of their daughters.</p>
<p><said direct="true">"My dear Mr. <name type="person">Bennet</name>,"</said> said his lady to him one day,
<said direct="true">"have you heard that Netherfield Park is let at last?"</said></p>
<p>Mr. <name type="person">Bennet</name> replied that he had not.</p>
<p><said direct="true">"But it is,"</said> returned she; <said direct="true">"for Mrs. <name type="person">
Long</name> has just been here, and she told me all about it."</said></p>
<p>Mr. <name type="person">Bennet</name> made no answer.</p>
<p><said direct="true">"Do you not want to know who has taken it?"</said> cried his wife impatiently.</p>
<p><said direct="true">"You want to tell me, and I have no objection to hearing it."</said></p>
<p>This was invitation enough.</p>
```

Figure 1: An extract showing the encoding of functional metadata

In Figure 2 an example of the encoding of the morphological, syntactic and semantic metadata is given. The part-of-speech category is encoded in the pos attribute. The meaning (sense) is encoded in the WordNet attribute. The grammar and relationships (e.g., subject) between the words are encoded in the dependencies tags.



```
<tokens>
    <w id="1" pos="PRP">I</w>
    <w id="2" pos="VBP" wordnet="know.v.01">know</w>
    <w id="3" pos="DT">all</w>
    <w id="4" pos="PRP$">my</w>
    <w id="5" pos="NN" wordnet="property.n.01">property</w>
    <w id="6" pos=",">,</w>
    <w id="7" pos="CC">and</w>
    <w id="8" pos="WRB">where</w>
    <w id="9" pos="DT">the</w>
    <w id="10" pos="NN" wordnet="money.n.01">money</w>
    ...
</tokens>
<dependencies type="basic-dependencies">
    <dep type="root">
        <governor idx="0">ROOT</governor>
        <dependent idx="2">know</dependent>
    </dep>
    <dep type="nsubj">
        <governor idx="2">know</governor>
        <dependent idx="1">I</dependent>
    </dep>
```

Figure 2: An extract showing the encoding of morphological, syntactic and semantic metadata

The next step was to test the ability of a system to process the texts with these metadata and allow a user to retrieve specific instances.

Development of the Prototype

Various technologies were used in the development of this prototype, particularly, Node.js, MySQL, JavaScript, HTML (hypertext markup language) and CSS (cascading style sheets). Node.js is a JavaScript runtime environment (https://nodejs.dev/). The back end (i.e., processing and algorithms) of the system was written in Node.js, whereas the database was a MySQL database (https://www.mysql.com/). The front end (interface) was written in JavaScript, HTML and CSS.

The two encoded files for each text are processed and parsed by the prototype. The data are then stored in a relational database. It is faster and more efficient to query a relational database than to query XML (extensible markup language) files. Furthermore, as the encoded metadata are in two files, it would complicate the query process.

The bibliographic details from the one file are taken to create a unique document in the database. The system then takes the sentences from the other file to add the words with the morphological, syntactic and semantic information to the database. A check is done in order to ensure the words are not duplicated in the database, but all relevant words are still linked to the specific text. The system then moves back to the first file to process the functional encoding. To include this information in the database, a structure that indicates the relationships between words is formed, so that it is known if a word is part of direct speech, which is part of a paragraph, which is part of the body of the text, etc. More detail about the development and working of the prototype can be found in Ball (2020).

The next section will illustrate the advanced retrieval that is possible on this prototype.

Results

The prototype, called *inkling*, has a simple interface that is sufficient to test functionality. As was explained in the methodology, the purpose of the prototype is not to be ready for production, but to test a concept. The way in which the prototype enables advanced retrieval will be explained.

The home page (see Figure 3) has a search bar at the top, filtering options in a panel to the left and an interactive area in which the user can build a search. The search bar is not functional in the prototype. The search functionality is tested using the graphical interface below the search bar. The search bar is included to show how a possible advanced query language can be used for querying the database, but the implementation was not part of the prototype.

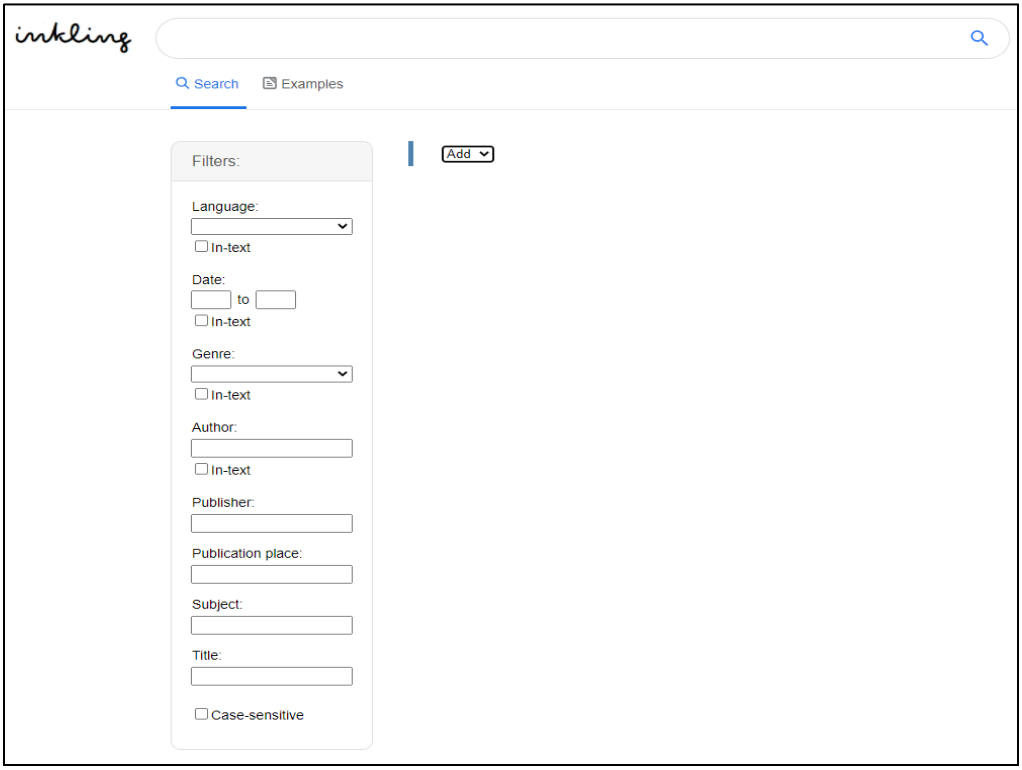
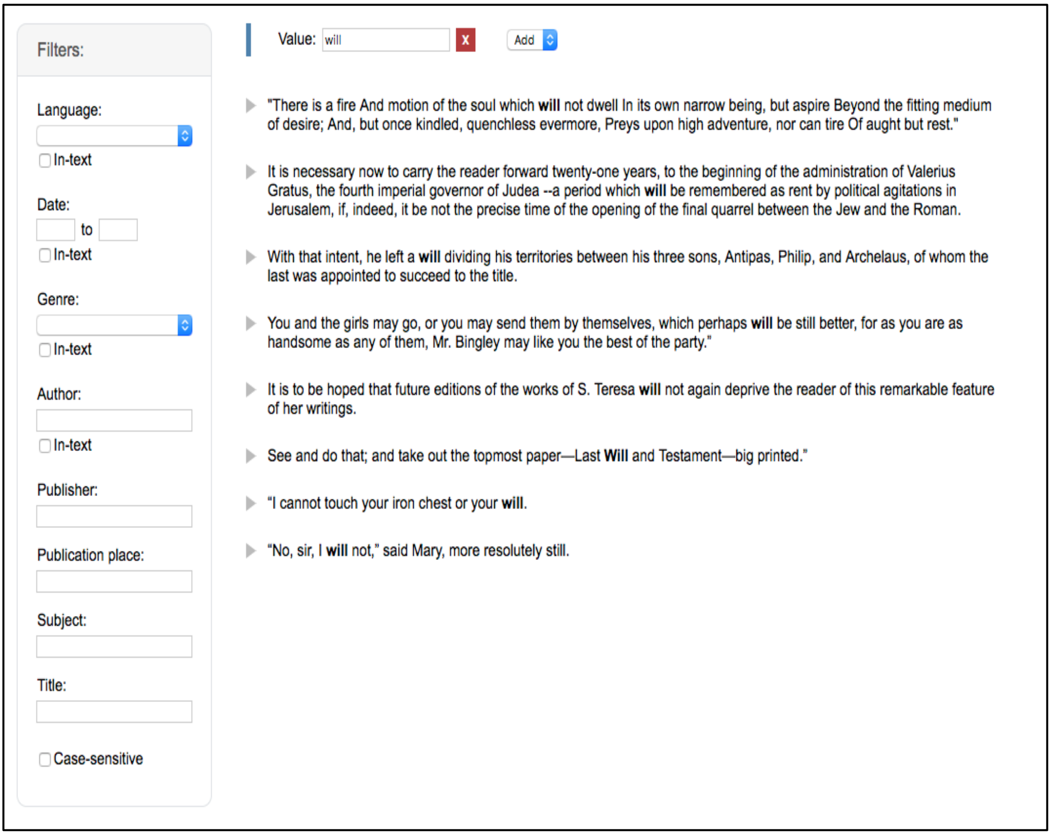


Figure 3: The home page of the prototype, *inkling*

The interface allows a user to construct a search by adding more items. The simplest search is to search for a value. Figure 4 shows a simple search for all instances of “will.” The results of the search are displayed below the search bar. The value that was searched for is highlighted and some context is given to the search. It is possible to expand each instance to see more context and to see some bibliographic metadata (e.g., title of the text, the author and publication date). By using the information from the encoded metadata, it is possible to create a complex search. **Error! Reference source not found.**Figure 5 shows a complex search where properties from the morphological level, functional level and bibliographic level are combined. Only instances of “will” that are nouns and that appear in direct speech in texts written by the author Eliot are returned.



Value: will

▶ "There is a fire And motion of the soul which **will** not dwell In its own narrow being, but aspire Beyond the fitting medium of desire; And, but once kindled, quenchless evermore, Preys upon high adventure, nor can tire Of aught but rest."

▶ It is necessary now to carry the reader forward twenty-one years, to the beginning of the administration of Valerius Gratus, the fourth imperial governor of Judea --a period which **will** be remembered as rent by political agitations in Jerusalem, if, indeed, it be not the precise time of the opening of the final quarrel between the Jew and the Roman.

▶ With that intent, he left a **will** dividing his territories between his three sons, Antipas, Philip, and Archelaus, of whom the last was appointed to succeed to the title.

▶ You and the girls may go, or you may send them by themselves, which perhaps **will** be still better, for as you are as handsome as any of them, Mr. Bingley may like you the best of the party."

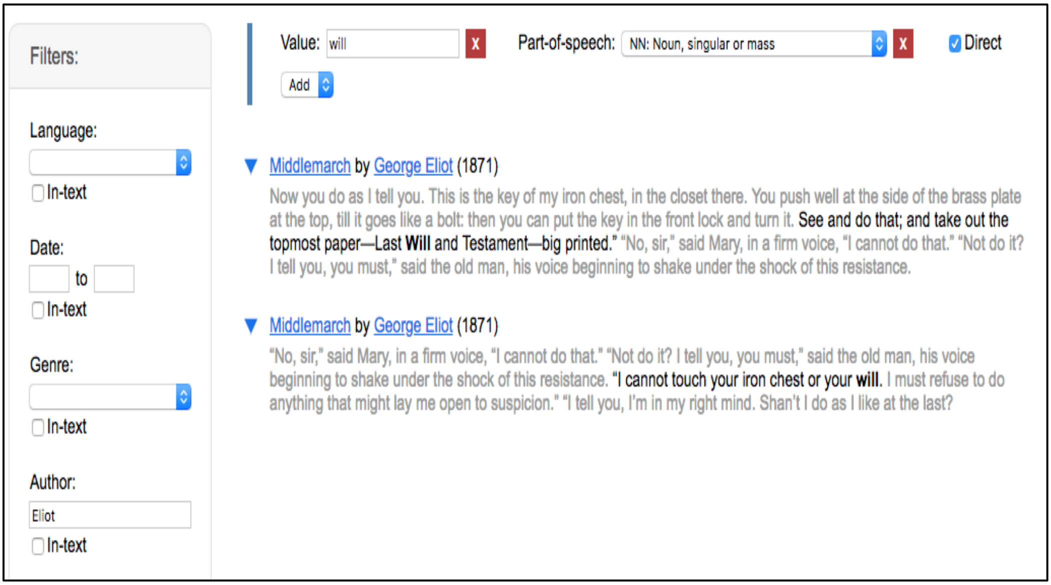
▶ It is to be hoped that future editions of the works of S. Teresa **will** not again deprive the reader of this remarkable feature of her writings.

▶ See and do that; and take out the topmost paper—Last **Will** and Testament—big printed."

▶ "I cannot touch your iron chest or your **will**.

▶ "No, sir, I **will** not," said Mary, more resolutely still.

Figure 4: Searching for “will”



Filters:

Language:

□ In-text

Date: to

□ In-text

Genre:

□ In-text

Author: Eliot

□ In-text

Value: will X Part-of-speech: NN: Noun, singular or mass X ☑ Direct

Add

▼ Middlemarch by George Eliot (1871)

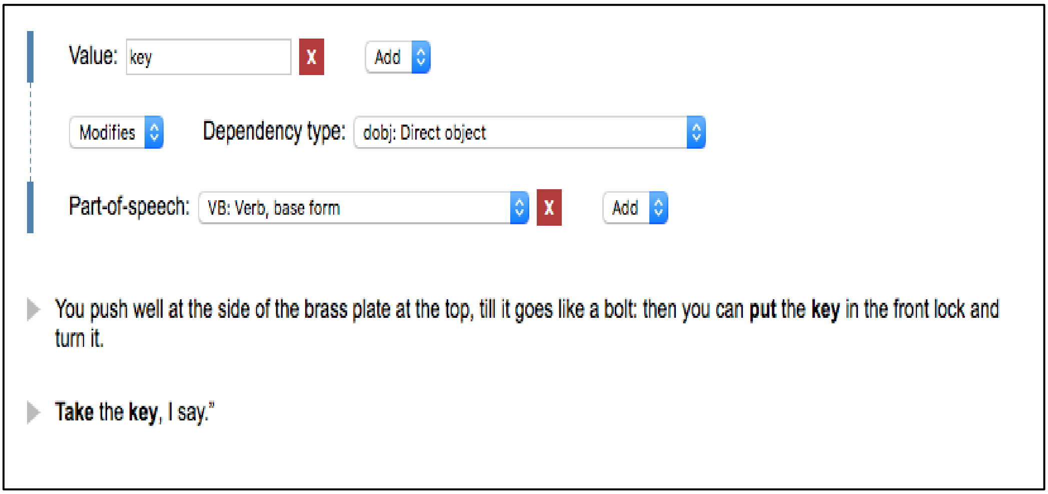
Now you do as I tell you. This is the key of my iron chest, in the closet there. You push well at the side of the brass plate at the top, till it goes like a bolt: then you can put the key in the front lock and turn it. **See and do that; and take out the topmost paper—Last Will and Testament—big printed.**" "No, sir," said Mary, in a firm voice, "I cannot do that." "Not do it? I tell you, you must," said the old man, his voice beginning to shake under the shock of this resistance.

▼ Middlemarch by George Eliot (1871)

"No, sir," said Mary, in a firm voice, "I cannot do that." "Not do it? I tell you, you must," said the old man, his voice beginning to shake under the shock of this resistance. **"I cannot touch your iron chest or your will.** I must refuse to do anything that might lay me open to suspicion." "I tell you, I'm in my right mind. Shan't I do as I like at the last?

Figure 5: Searching for "will," where it is a noun, in direct speech and the author is Eliot

Searching according to the syntactic level is demonstrated in Figure 6 where the direct object is "key" and it is used in combination with a verb.



Value: key X Add

Modifies Dependency type: dobj: Direct object

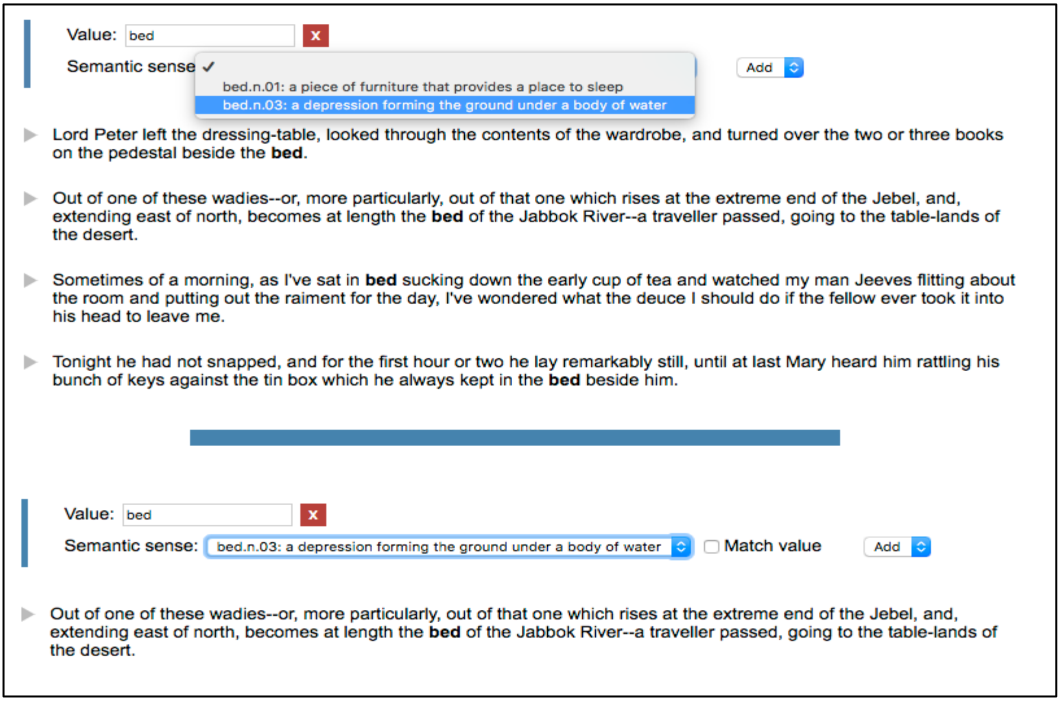
Part-of-speech: VB: Verb, base form X Add

You push well at the side of the brass plate at the top, till it goes like a bolt: then you can **put** the **key** in the front lock and turn it.

Take the **key**, I say."

Figure 6: Searching for instances where "key" is the direct object

Searching using the semantic metadata is illustrated in Figure 7. This figure consists of two parts. First, all the instances of “bed” are returned. The user then specifies that the instances of “bed” must refer to the bottom of a lake or river. Once the filter for the specific sense has been selected, only instances with the relevant sense are returned.


✓
bed.n.01: a piece of furniture that provides a place to sleep
bed.n.03: a depression forming the ground under a body of water
Add
x
Add
x
□ Match value

Value: bed

Semantic sense

▶ Lord Peter left the dressing-table, looked through the contents of the wardrobe, and turned over the two or three books on the pedestal beside the **bed**.

▶ Out of one of these wadies--or, more particularly, out of that one which rises at the extreme end of the Jebel, and, extending east of north, becomes at length the **bed** of the Jabbok River--a traveller passed, going to the table-lands of the desert.

▶ Sometimes of a morning, as I've sat in **bed** sucking down the early cup of tea and watched my man Jeeves flitting about the room and putting out the raiment for the day, I've wondered what the deuce I should do if the fellow ever took it into his head to leave me.

▶ Tonight he had not snapped, and for the first hour or two he lay remarkably still, until at last Mary heard him rattling his bunch of keys against the tin box which he always kept in the **bed** beside him.

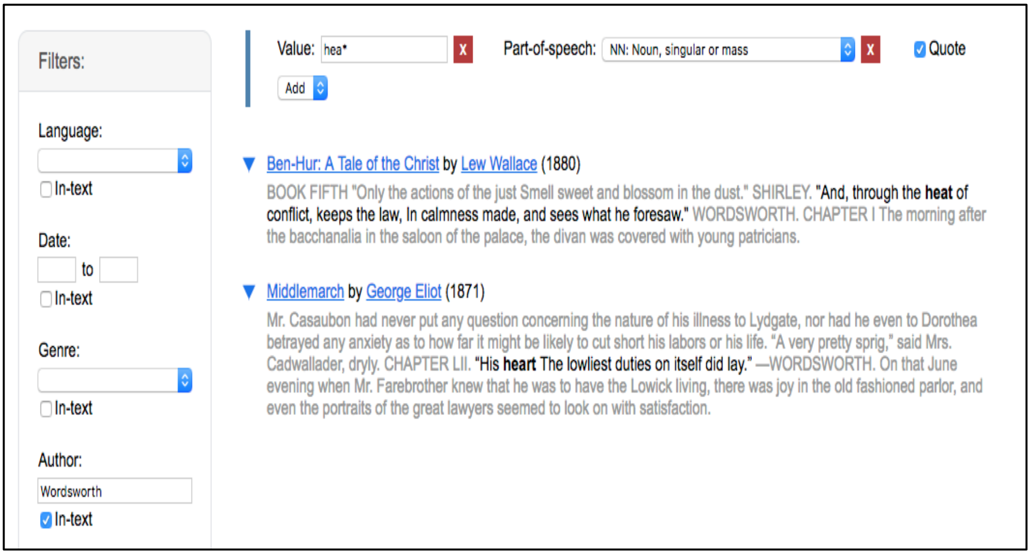
Value: bed

Semantic sense: bed.n.03: a depression forming the ground under a body of water

▶ Out of one of these wadies--or, more particularly, out of that one which rises at the extreme end of the Jebel, and, extending east of north, becomes at length the **bed** of the Jabbok River--a traveller passed, going to the table-lands of the desert.

Figure 7: Using semantic metadata in a search

Some other search features on this prototype are the ability to search for truncated forms of words, words near other words (proximity) and inflected forms. Figure 8 demonstrates some of these advanced features, as well as the value of in-text bibliographic metadata. Truncation is used to search for all instances that start with *hea*-. Further filters are applied, and the search is for items that are nouns, that appear in quoted texts, where the author of the quotations is Wordsworth. Two quotations by Wordsworth that include a noun that starts with *hea*- are found in the database. The first is a quote in *Ben-Hur* by Lew Wallace that includes the word “heat,” and the second is a quotation in *Middlemarch* by George Eliot that includes the word “heart.” The in-text metadata is used to find words by an author that is not the same as the author of the volume.



Value: hea*

Part-of-speech: NN: Noun, singular or mass

☑ Quote

Author: Wordsworth

☑ In-text

▼ Ben-Hur: A Tale of the Christ by Lew Wallace (1880)

BOOK FIFTH "Only the actions of the just Smell sweet and blossom in the dust." SHIRLEY. "And, through the **heat** of conflict, keeps the law, In calmness made, and sees what he foresaw." WORDSWORTH. CHAPTER I The morning after the bacchanalia in the saloon of the palace, the divan was covered with young patricians.

▼ Middlemarch by George Eliot (1871)

Mr. Casaubon had never put any question concerning the nature of his illness to Lydgate, nor had he even to Dorothea betrayed any anxiety as to how far it might be likely to cut short his labors or his life. "A very pretty sprig," said Mrs. Cadwallader, dryly. CHAPTER LII. "His **heart** The lowliest duties on itself did lay." —WORDSWORTH. On that June evening when Mr. Farebrother knew that he was to have the Lowick living, there was joy in the old fashioned parlor, and even the portraits of the great lawyers seemed to look on with satisfaction.

Figure 8: Searching in quoted text

Another example (not demonstrated here) of in-text bibliographic metadata is the use of language. A search for instances that end in -*les* in English texts could return instances that are in other languages, as the language of the text (volume) is recorded as English. This means quotations in other language can be included. However, if the user specifies that the in-text language should be English, then the phrases in foreign languages are excluded.

Discussion

The prototype discussed in this article demonstrates that fine-grained retrieval is possible if texts are encoded with detailed metadata. The metadata expose properties of the text and words in the texts so that they can be used for retrieval. Some properties of texts and words can be easily identified by humans. For example, a human should be able to see easily if something is a heading, where one paragraph starts and ends or what a word means in a certain context. This information needs to be made explicit to the computer. This can be done by encoding texts with granular metadata. In this prototype the encoded texts were parsed, and the information stored in a database. The database then contained detailed information about each word in a text and fine-grained retrieval is made possible.

The fine-grained retrieval was demonstrated through the examples. It was shown that it is possible to search using different categories of metadata, as well as to combine categories of metadata to do very precise retrieval. For example, it is possible to search for a word with a specific part-of-speech category, with a specific meaning, in direct speech and written by a certain author. The in-text bibliographic metadata also enables exact retrieval, making it possible to drill down into texts and extract only that which the researcher is interested in, such as quotations from Shakespeare used in nineteenthcentury publications, or selecting essays from authors included in a single volume (as discussed in the introduction).

The value of encoding texts with granular metadata is clear. However, the process is time-consuming and resource intensive. As such, the use of artificial intelligence (AI) should be explored to see to what extent the encoding can be automated. This study already relied on software to assist with the some of the encoding (morphological and syntactic categories). Although there are no programs that are reliable enough currently to automate the encoding of the other categories of metadata, there are developments in these fields (e.g., Ustalov et al. 2018).

This study noted the limitations of tools that explore digital text collections where no detailed filtering is possible. If users view trends of certain words over a period of time, but the collection includes anthologies or compendiums or works with quotations from other authors, then the results are mixed and not entirely accurate. It is important to consider the size of the collection or dataset and the purpose of the research. One could argue that if words from quotations from other authors or words that are in a different language are retrieved from a very large collection, it will not have a significant effect on the statistics, and one could still observe trends. However, the value of smaller well-curated datasets that enable precise retrieval should be considered.

Furthermore, the contribution of different role players to enable more effective retrieval and the role information professionals can play should be considered. Information professionals should contribute to this fast-changing landscape of growing data and ways to explore it. Cox (2021) suggests that information services can support users who wish to analyse content in new ways, for example, by providing support, content and infrastructure. Senseney et al. (2021) explore the role of librarians and information professionals in text and data mining, explaining that digital scholarship tools and methods are being integrated in library services and that librarians have a role to play to facilitate data-intensive research. This study has shown the value that granular metadata can add. Information professionals could contribute here.

Yet, this is an interdisciplinary task. The metadata fields that are required to meet the needs of researchers should come from intensive user-needs studies by researchers, particularly in the humanities. Linguists should contribute to the way in which words and properties of words can be described. Computer scientists should contribute to the development of automated encoding software and infrastructure. User experience (UX) experts should contribute to the design of interfaces. Information professionals should contribute to bibliographic metadata and other categories of metadata, as well as be involved in extensive information-seeking experiments and evaluating the effectiveness and efficiency of the systems. Furthermore, information professionals could act as the force that brings the different role players together.

Future work could look at the development of a fully-fledged retrieval system and the scalability of the database, the automation of the encoding of different categories of metadata, the development of interfaces that are user-friendly and can accommodate the needs of laypersons and experts, as well as the identification of metadata fields for different user groups.

Conclusion

The amount of digital information is growing. Technology offers different ways in which users can engage with digital text collections. Though much research and developments have been done, this study demonstrated that using granular metadata to encode texts, very precise retrieval is made possible, and users can select exactly the instances that they need and exclude that which is irrelevant. In a world where the amount of information is increasing and users need to be able to answer questions quickly and accurately, more should be done to enable precise retrieval.

References

Ball, Liezl H. 2020. “Enhancing Digital Text Collections with Detailed Metadata to Improve Retrieval.” PhD diss., University of Pretoria. http://hdl.handle.net/2263/79015

Ball, Liezl H., and Theo J. D. Bothma. 2022. “Investigating the Extent to Which Words or Phrases with Specific Attributes Can Be Retrieved from Digital Text Collections.” *Information Research* 27 (1): 917. https://doi.org/10.47989/irpaper917

Cox, Andrew M. 2021. *Research Report: The Impact of AI, Machine Learning, Automation and Robotics on the Information Professions*. CILIP (The Library and Information Association). Accessed April 27, 2022. https://www.cilip.org.uk/page/researchreport

Edmond, Jennifer, and Jörg Lehmann. 2021. “Digital Humanities, Knowledge Complexity, and the Five ‘Aporias’ of Digital Research.” *Digital Scholarship in the Humanities* 36 (2): ii95–ii108. https://doi.org/https://doi.org/10.1093/llc/fqab031

Fenlon, Katrina, Megan Senseney, Harriett Green, Sayan Bhattacharyya, Craig Willis, and J. Stephen Downie. 2014. “Scholar-Built Collections: A Study of User Requirements for Research in Large-Scale Digital Libraries.” *Proceedings of the American Society for Information Science and Technology* 51 (1): 1–10. https://doi.org/https://doi.org/10.1002/meet.2014.14505101047

Finlayson, Mark A. 2015. “*ProppLearner*: Deeply Annotating a Corpus of Russian Folktales to Enable the Machine Learning of a Russian Formalist Theory.” *Digital Scholarship in the Humanities* 32 (2): 284–300. https://doi.org/https://doi.org/10.1093/llc/fqv067

Google Books Ngram Viewer. n.d. “Google Books Ngram Viewer Info.” Accessed August 18, 2020. https://books.google.com/ngrams/info

Heiden, Serge. 2010. “The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme.” In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC24)*, edited by Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto and Yasunari Harada, 389–398. Sendai: Institute for Digital Enhancement of Cognitive Development, Waseda University. https://aclanthology.org/Y10-1044/

Heuser, Ryan, Franco Moretti, and Erik Steiner. 2016. “The Emotions of London.” Pamphlets of the Stanford Literary Lab, Pamphlet 13. Accessed August 2, 2018. https://litlab.stanford.edu/LiteraryLabPamphlet13.pdf

Hoffmann, Sebastian, and Stefan Evert. 2006. “BNC*web* (CQP-edition): The Marriage of Two Corpus Tools.” In *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, Vol. 3, edited by Sabine Braun, Kurt Kohn and Joybrato Mukherjee, 177–195. Frankfurt: Peter Lang.

Jett, Jacob, Terhi Nurmikko-Fuller, Timothy W. Cole, Kevin R. Page, and J. Stephen Downie. 2016. “Enhancing Scholarly Use of Digital Libraries: A Comparative Survey and Review of Bibliographic Metadata Ontologies.” In *JCDL ’16: Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries*, 35–45. New York: The Association for Computing Machinery. https://doi.org/10.1145/2910896.2910903

Klimczak, Erik. 2013. *Design for Software: A Playbook for Developers*. Chichester: John Wiley and Sons.

Lansdall-Welfare, Thomas, and Nello Cristianini. 2020. “History Playground: A Tool for Discovering Temporal Trends in Massive Textual Corpora.” *Digital Scholarship in the Humanities* 35 (2): 328–341. https://doi.org/https://doi.org/10.1093/llc/fqy077

Lin, Yuri, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. “Syntactic Annotations for the Google Books Ngram Corpus.” In *ACL 2012: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 169–174. Stroudsburg, PA: Association for Computational Linguistics. https://aclanthology.org/P12-3029.pdf

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, the Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2010. “Quantitative Analysis of Culture Using Millions of Digitized Books.” *Science* 331 (6014): 176–182. https://doi.org/10.1126/science.1199644

Nguyen, Dong, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2020. “How We Do Things with Words: Analyzing Text as Social and Cultural Data.” *Frontiers in Artificial Intelligence* 3: 62. https://doi.org/10.3389/frai.2020.00062

Senseney, Megan, Eleanor Dickson Koehl, Beth Sandor Namachchivaya, and Bertram Ludäscher. 2021. *Transforming Library Services for Computational Research with Text Data: Environmental Scan, Stakeholder Perspectives, and Recommendations for Libraries*. Chicago: Association of College and Research Libraries. Accessed April 27. 2022. https://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/TransformingLibServices.pdf

Suranto, Beni. 2015. “Software Prototypes: Enhancing the Quality of Requirements Engineering Process.” In *Proceedings of ISTMET 2015 2nd International Symposium on Technology Management and Emerging Technologies*, 148–153. Piscataway: Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ISTMET.2015.7359019

TEI (Text Encoding Initiative). n.d. “TEI: Text Encoding Initiative.” Accessed January 12, 2018. http://www.tei-c.org/index.xml

Underwood, Ted. 2015. “Understanding Genre in a Collection of a Million Volumes.” White Paper Report 109365, University of Illinois, Urbana-Champaign. Accessed July 30, 2019. https://hcommons.org/deposits/item/hc:12277/

Underwood, Ted, David Bamman, and Sabrina Lee. 2018. “The Transformation of Gender in English-Language Fiction.” *Journal of Cultural Analytics* 3 (2): 1–25. https://doi.org/10.22148/16.019

Ustalov, Dmitry, Denis Teslenko, Alexander Panchenko, Mikhail Chernoskutov, Chris Biemann, and Simone Paolo Ponzetto. 2018. “An Unsupervised Word Sense Disambiguation System for Under-Resourced Languages.” In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, edited by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis and Takenobu Tokunaga, 1018–1022. Miyazaki: European Language Resources Association. https://aclanthology.org/L18-1164

Viiri, Sampo. 2014. *Digital Humanities and Future Archives*. London: Finnish Institute in London. Accessed September 29, 2020. https://www.fininst.uk/wp-content/uploads/2017/09/Digital_Humanities_and_Future_Archives.pdf

Walker, Miriam, Leila Takayama, and James A. Landay. 2002. “High-Fidelity or Low-Fidelity, Paper or Computer? Choosing Attributes When Testing Web Prototypes.” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 46 (5): 661–665. https://doi.org/10.1177/154193120204600513

Welsh, Megan E. 2014. “Review of Voyant Tools.” *Collaborative Librarianship* 6 (2): 96–98.